Prediction of Mortality from Respiratory Distress Among Long-Term Mechanically Ventilated Patients

Gregory Boverman[1] and Sahika Genc[1]

Abstract— **With the advent of inexpensive storage, pervasive networking, and wireless devices, it is now possible to store a large proportion of the medical data that is collected in the intensive care unit (ICU). These data sets can be used as valuable resources for developing and validating predictive analytics. In this report, we focus on the problem of prediction of mortality from respiratory distress among long-term mechanically ventilated patients using data from the publicly-available MIMIC-II database. Rather than only reporting p-values for univariate or multivariate regression, as in previous work, we seek to generate sparsest possible model that will predict mortality. We find that the presence of severe sepsis is highly associated with mortality. We also find that variables related to respiration rate have more predictive accuracy than variables related to oxygenation status. Ultimately, we have developed a model which predicts mortality from respiratory distress in the ICU with a cross-validated area-under-the-curve (AUC) of approximately 0.74. Four methodologies are utilized for model dimensionality-reduction: univariate logistic regression, multivariate logistic regression, decision trees, and penalized logistic regression.**

I. INTRODUCTION

The term "acute respiratory distress syndrome", or ARDS, has been critized as being somewhat lacking in objective validity and predictive accuracy [**?**]. The original definition was promulgated by the American-European Consensus Conference (AECC) in 1994 [**?**]. ARDS was at that time defined as a disorder with acute onset, poor oxygenation as evidenced by an arterial PaO_2/FiO_2 ratio of less than 200 mm Hg, the presence of bilateral infiltrates as seen on a chest radiograph, and pulmonary artery wedge pressure of less than 18 mm Hg, with no evidence of left atrial hypertension. The definition was recently revised, in 2012, with many of the limitations of the original AECC definition addressed [**?**]. In addition to the original oxygenation variable, several other variables were added, namely C_{RS}, the respiratory system compliance, and the positive end-expiratory pressure (PEEP) applied by the artificial respirator to maintain airway clearance. However, even the revised definition only resulted in an area under the receiver operating curve of 0.577 for prediction of mortality [**?**].

This report will focus on the prediction of mortality due to respiratory distress in ICU patients ventilated for longer than 48 hours. More specifically, we utilize information from the first 24 hours after the onset of mechanical ventilation to predict mortality. We objectively rank the available parameters and develop predictive models of mortality in an entirely data-driven manner, with the emphasis of the investigations being on the selection of the most relevant subset of the moderately high-dimensional feature set that is sufficient to predict the outcome.

This work makes use of MIMIC-II [**?**], a large comprehensive database of ICU records collected over a period of years at a Beth Israel Deaconess Hospital, a tertiary-care hospital affiliated with Harvard Medical School, and post-processed and de-identified by researchers at the Massachusetts Institute of Technology (MIT). Our work in this report in many ways follows the lead of research conducted by Jia [**?**] at MIT in 2007. We have extended Jia's investigations to the realm of machine learning, looking at the prediction of mortality from parameters measured in the initial stages of mechanical ventilation.

II. THE PREDICTION PROBLEM

The problem that we have set for ourselves is, given knowledge of the patient and ventilator state in the first 24 hours of mechanical ventilation, can we predict mortality ?

Even though we have a large amount of data with which to work, we also have a relatively large number of potential features, and it is necessary to intelligently select those features which are most informative, as, when the number of features grows, correlations will tend to appear between variables by chance alone.

A. Methodology

We began by parsing all patient data records within the MIMIC-II ICU database. Each ICU admission was regarded as its own independent event and death during this admission was defined by the reported date of death falling between between the reported start and end dates of the admission. As in the work by Jia, we ruled-out hospital admissions where the ICD-9 code for congestive heart failure was present. We further required a continuous mechanical ventilation time of 48 hours or greater and discarded records not containing at least one instance of each variable utilized in the classification. Ultimately, we were left with 1054 ICU admissions. Of these, there were 182 ICU admissions leading to mortality. Respiratory-distress-related mortality was defined as an ICU admission leading to mortality where the final PaO_2/FiO_2 ratio was less than 200 mm Hg. The number of admissions with the ICD-9 code for severe sepsis was 119, and the number with the ICD-9 code for ARDS was 258.

B. Classification approaches

In what follows, we describe quantitative classification results using a number of classification and feature-selection

[1] G. Boverman and S. Genc are with the GE Global Research Center, 1 Research Circle, Niskayuna, NY, 12309 USA `boverman at ge.com`

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

approaches, namely: univariate logistic regression, multivariate logistic regression, decision trees, and penalized logistic regression. We have chosen these approaches mainly because they produce classifiers which are in some sense "explainable".

All of the data analysis was performed using the R language [?]. The univariate and multivariate logistic regression approaches were implemented using the R "glm" package. The decision tree model was implemented using the R "party" package [?], and the penalized logistic regression approach utilized the R "penalized" package [?]. Analysis of receiver operating characteristic (ROC) curves was performed using the R "ROCR" package [?].

1) Univariate logistic regression: In the first instance, we utilized univariate logistic regression to determine which of the measured parameters were correlated with the outcome to the level of statistical significance, defined here as a p-value of 0.05 or below.

In table I, we show the results of univariate logistic regression of respiratory-distress-related mortality vs. the patient's state in the first 24 hours after initiation of mechanical ventilation. For each parameter, we display the standardized logistic regression coefficient (increase/decrease in risk vs. number of standard deviations from the mean) and the p-value. Additionally, we generated univariate classifiers for each parameter and reported the area under the receiver operating characteristics (ROC) curve for each of the classifiers using five-fold cross validation.

2) Multivariate logistic regression: We next turn our attention to multivariate logistic regression, first generating a model using all of the features simultaneously. This model achieved a cross-validated AUC of approximately 0.71. For this model, relatively few parameters achieved a significance level of 0.05. These parameters were, specifically: the presence of the ICD-9 code for sepsis, with a very high significance level and a coefficient value of 1.56, the age at admission, the minute volume in the second 12 hours of mechanical ventilation, and the respiration rate divided by the tidal volume in the second 12 hours of mechanical ventilation.

We then built a reduced-feature logistic regression model using only these four parameters, reasoning that the additional non-statistically-significant features were serving only as confounding factors. This model achieved an improved AUC of 0.74 The p-values and coefficients for the reduced-feature model are given in Table II. We see that the three non-binary variables are assigned roughly equal standardized weights.

3) Decision Trees: An alternative approach to feature-selection involves the use of decision trees [?], [?]. In the most common implementation of the decision tree, hard (typically binary) decisions are made at each node based on only a single variable. We have chosen the recursive splitting criterion of only allowing splits which achieve a significance level of 0.05. Fig. 1 shows the generated decision tree with the specified splitting rule. For the decision tree model, the cross-validated AUC was approximately 0.68, though with a higher variance than the logistic regression model, and the 95% CI was 0.68 +/- 0.035. Thus, the decision tree somewhat underperformed the multivariate logistic regression model.

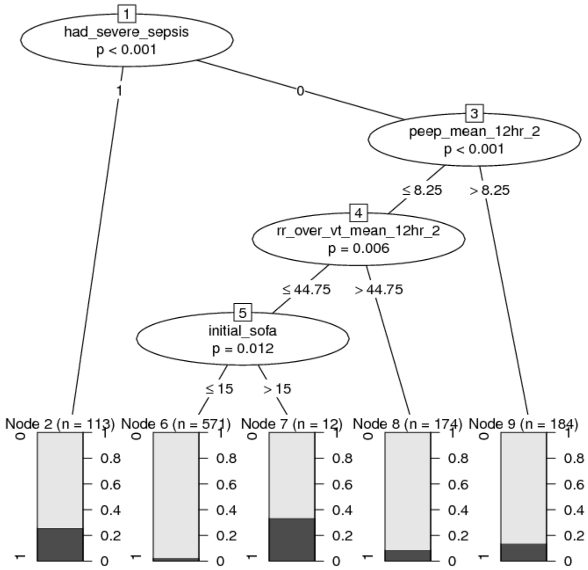

Fig. 1. Decision tree for prediction of respiratory-distress-related mortality.

The decision tree utilized two of the four parameters chosen by the multivariate logistic regression approach, namely the presence of the ICD-9 code for severe sepsis and the respiration rate divided by the tidal volume.

4) Penalized Logistic Regression: An interesting approach to feature-selection and solving the problem of over-fitting involves the addition of regularization functions to the functions being optimized by the logistic regression. Early work on penalized logistic regression utilized an l_2 penalty function, but somewhat more recently an l_1 penalty function, the "lasso" [?], has been introduced to encourage "sparsity" of the coefficient set.

An appealing aspect of the penalized regression method with an l_1 penalty is that we can visualize the relative values of the features by displaying their estimated standardized coefficient values as the l_1 regularization parameter varies.

We optimize the l_1 and l_2 regularization parameters by a "coordinate-descent" method, varying one parameter at a time, with the goal of maximizing the cross-validated log-likelihood. The regularization paths for all parameters are visualized in Fig. 2. We see that, in general, the presence of the ICD-9 code for severe sepsis is most highly predictive of respiratory-distress-related mortality. However, a number of other parameters, such as the initial SOFA score, persist over a greater range of regularization parameters. The estimated coefficients for the model utilizing the optimal values of the regularization parameters are given in Table III. The mean cross-validated AUC was 0.702, with a 95% CI of 0.037.

III. Conclusion

There has been debate in the clinical community whether the term "ARDS" has sufficient clinical validity. In our data-

TABLE I

UNIVARIATE REGRESSION OF RESPIRATORY-DISTRESS-RELATED MORTALITY

Parameter	Description	Coefficient	p-value	Cross-validated AUC
Subject male ?	Was the subject male	-0.135	0.551	0.52
ICU Admission	Admission number	-0.18	0.532	0.51
Admission age	Age at admission	0.154	0.125	0.55
Initial weight	Weight at admission	-0.17	0.164	0.54
Days ventilated	Number of days ventilated	0.11	0.248	0.59
Initial SOFA	SOFA score at admission	0.62	4.56e-8*	0.66
Initial SAPSI	SAPS-I score at admission	0.44	8.89e-5*	0.61
FiO$_2$ 12 hr 1	FiO$_2$ mean in first 12 hours	0.32	0.0027*	0.59
FiO$_2$ 12 hr 2	FiO$_2$ mean in second 12 hours	0.37	9.54e-5*	0.60
severe sepsis	ICD-9 code for severe sepsis present ?	1.68	4.77e-11*	0.61
ARDS	ICD-9 code for ARDS present ?	0.61	0.01*	0.56
V$_t$ 12 hr 1	V$_t$ mean in first 12 hours	-0.26	0.0254*	0.58
V$_t$ 12 hr 2	V$_t$ mean in second 12 hours	-0.23	0.08	0.57
RR 12 hr 1	Respiration rate mean in first 12 hours	0.42	3.2e-5*	0.64
RR 12 hr 2	Respiration rate mean in second 12 hours	0.56	4.17e-8*	0.68
PIP 12 hr 1	Peak inspiratory pressure mean in first 12 hours	0.19	0.073	0.55
PIP 12 hr 2	Peak inspiratory pressure mean in second 12 hours	0.34	9.6e-4*	0.58
Minvol 12 hr 1	Minute volume mean in first 12 hours	0.26	0.0095*	0.59
Minvol 12 hr 2	Minute volume mean in second 12 hours	0.41	3.05e-5*	0.62
HR 12 hr 1	Heart rate mean in first 12 hours	0.11	0.30	0.51
HR 12 hr 2	Heart rate mean in second 12 hours	0.02	0.88	0.43
PEEP 12 hr 1	Positive End-Expiratory Pressure mean in first 12 hours	0.24	0.007*	0.59
PEEP 12 hr 2	Positive End-Expiratory Pressure mean in second 12 hours	0.35	7.8e-5*	0.61
SaO$_2$ 12 hr 1	arterial oxygen saturation mean, first 12 hours	-0.09	0.224	0.58
SaO$_2$ 12 hr 2	arterial oxygen saturation mean, second 12 hours	-0.11	0.158	0.57
Systbp 12 hr 1	systolic blood pressure mean, first 12 hours	-0.35	0.002*	0.60
Systbp 12 hr 2	systolic blood pressure mean, second 12 hours	-0.36	0.0008*	0.59
PaO$_2$ 12 hr 1	blood oxygenation mean, first 12 hours	-0.24	0.055	0.59
PaO$_2$ 12 hr 2	blood oxygenation mean, second 12 hours	-0.37	0.012*	0.61
Plat 12 hr 1	Plateau pressure mean, first 12 hours	0.28	0.0046*	0.58
Plat 12 hr 2	Plateau pressure mean, first 12 hours	0.40	5.85e-5*	0.59
CRS 12 hr 1	Respiratory compliance mean, first 12 hours	-0.24	0.051	0.57
CRS 12 hr 2	Respiratory compliance mean, second 12 hours	-0.26	0.047*	0.60
pH 12 hr 1	pH mean, first 12 hours	-0.47	6.9e-6*	0.63
pH 12 hr 2	pH mean, second 12 hours	-0.44	3.42e-5*	0.61
V$_t$ normalized 12 hr 1	Normalized V$_t$ mean in first 12 hours	-0.034	0.77	0.48
V$_t$ normalized 12 hr 2	Normalized V$_t$ mean in second 12 hours	-0.038	0.74	0.48
RR/V$_t$ 12 hr 1	RR/V$_t$ mean in first 12 hours	0.35	2.4e-4*	0.62
RR/V$_t$ 12 hr 2	RR/V$_t$ mean in second 12 hours	0.41	1.1e-4*	0.62
PaO$_2$/FiO$_2$ 12 hr 1	Normalized blood oxygenation mean, first 12 hours	-0.43	0.0016*	0.62
PaO$_2$/FiO$_2$ 12 hr 2	Normalized blood oxygenation mean, second 12 hours	-0.50	5.8e-4*	0.62

TABLE II

MULTIVARIATE LOGISTIC REGRESSION FOR PREDICTION OF MORTALITY WITH REDUCED FEATURE SET

Parameter	Description	Coefficient	p-value
Admission age	Age at admission	0.30	0.0027
severe sepsis	ICD-9 code for severe sepsis present	1.45	9.24e-8
Minvol 12 hr 2	Minute volume mean in second 12 hours	0.33	0.0029
RR/V$_t$ 12 hr 2	RR/V$_t$ mean in second 12 hours	0.31	0.0036

mining of the MIMIC-II ICU database, admittedly a data set with limited diversity, for example from the geographical perspective, we have found that the presence of the ICD-9 code for respiratory distress is not very predictive of mortality. However, we were able to find sets of parameters with high cross-validated prediction accuracy. Interestingly, variables related to oxygenation, for example the PaO$_2$/FiO$_2$ ratio, were not generally selected by the models that were implemented although the ratio appears as an ARDS-related mortality predictor. Thus, there is more work to better understand the various relationships among variables while testing different hypothesis. Finally, we have started to construct models to test various scenarios and consult with subject matter experts to determine the medical implications of our findings.

ACKNOWLEDGMENT

The authors acknowledge Kalpit Desai and Rohan Roy Choudhury for the use of their scripts for training and evaluation of classifiers in R.

TABLE III
Penalized logistic regression of respiratory-distress-related mortality

Parameter	Description	Coefficient
Admission age	Age at admission	0.18
Initial weight	Weight at admission	-0.19
Initial SOFA	SOFA score at admission	0.19
Initial SAPSI	SAPS-I score at admission	0.12
FiO_2 12 hr 2	FiO_2 mean in second 12 hours	0.07
severe sepsis	ICD-9 code for severe sepsis present ?	0.59
V_t 12 hr 1	V_t mean in first 12 hours	-0.04
Minvol 12 hr 2	Minute volume mean in second 12 hours	0.21
HR 12 hr 2	Heart rate mean in second 12 hours	-0.11
Systbp 12 hr 2	systolic blood pressure mean, second 12 hours	-0.08
PaO_2 12 hr 2	blood oxygenation mean, second 12 hours	-0.01
Plat 12 hr 2	Plateau pressure mean, first 12 hours	0.14
pH 12 hr 2	pH mean, second 12 hours	-0.11
RR/V_t 12 hr 2	RR/V_t mean in second 12 hours	0.23
PaO_2/FiO_2 12 hr 2	Normalized blood oxygenation mean, second 12 hours	-0.13

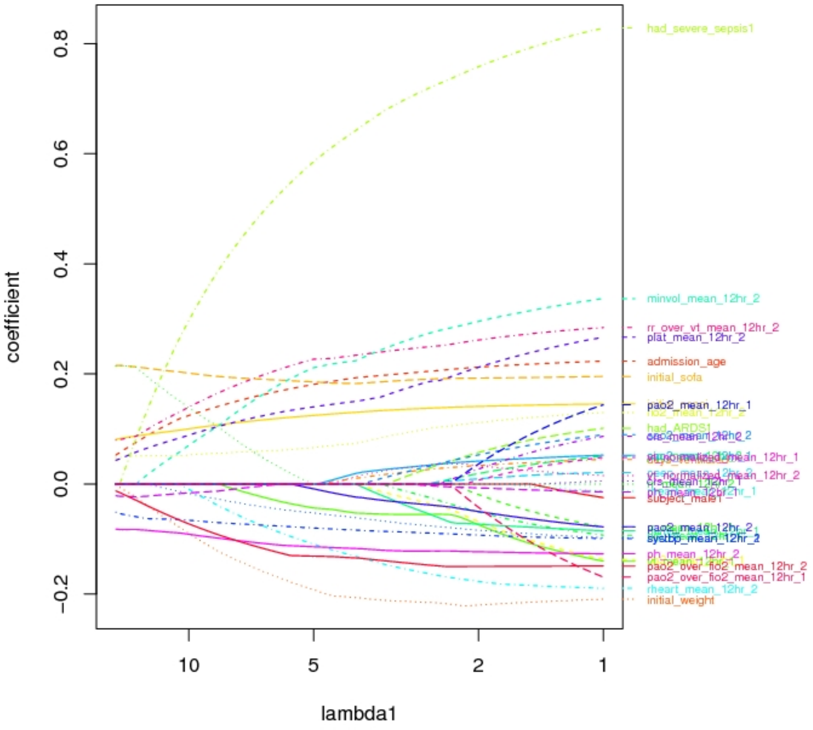

Fig. 2. Regularization paths of all features used in the classification vs. the l_1 regularization parameter